(19) JAPANESE PATENT OFFICE

PATENT ABSTRACTS OF JAPAN

(11) Publication number: **08050593 A**

(43) Date of publication of application: **20.02.96**

(51) Int. Cl **G06F 17/30**

(21) Application number: **06202832**

(22) Date of filing: **04.08.94**

(71) Applicant: **FUJI XEROX CO LTD**

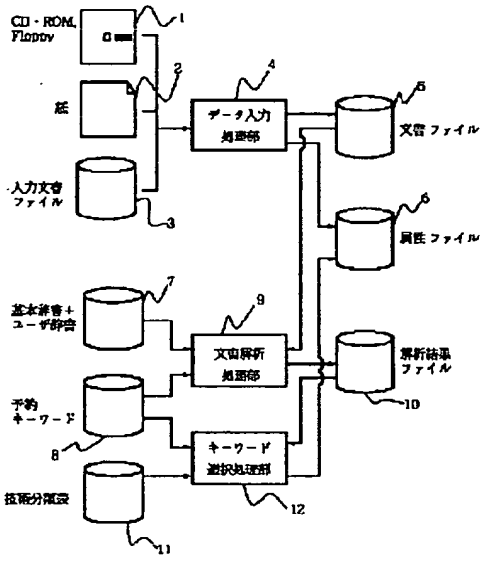
(72) Inventor: **YAMAGUCHI YOSHIAKI**

(54) KEYWORD IMPARTING DEVICE

(57) Abstract:

PURPOSE: To impart keywords to a document corresponding to the various states of using keywords and document databases based on the sense of a user.

CONSTITUTION: This device is provided with a means 9 for scanning the objective document and extracting the words, and a means 9 for counting the weight based on the prescribed reference of the extracted word, a means 11 for holding the reference data with keyword including the arrangement according to the prescribed reference of two or more related words corresponding to the candidate of the keyword to be imparted and the keyword candidate, a means 12 for collating the arrangement according to the prescribed reference of the extracted word with the arrangement of the related word of the keyword imparting reference data, and a means 12 for imparting the keyword candidate corresponding to the related word to which arrangements are matched with each other as the keyword of the objective document. The candidate of the set keyword is imparted as the document keyword based on the word extracted from the document.



(19)日本国特許庁（ＪＰ）　(12) **公開特許公報**（Ａ）　(11)特許出願公開番号

特開平8－50593

(43)公開日　平成８年(1996)２月20日

(51)Int.Cl.[6]	識別記号	庁内整理番号	ＦＩ		技術表示箇所
Ｇ０６Ｆ　17/30					
		9194－5L	Ｇ０６Ｆ　15/ 401	３１０　Ｃ	

審査請求　未請求　請求項の数１　ＦＤ　（全 14 頁）

(21)出願番号　特願平6－202832

(22)出願日　平成６年(1994)８月４日

(71)出願人　000005496
富士ゼロックス株式会社
東京都港区赤坂三丁目３番５号

(72)発明者　山口　義昭
神奈川県川崎市高津区坂戸３丁目２番１号
ＫＳＰ　Ｒ&Ｄ　ビジネスパークビル
富士ゼロックス株式会社内

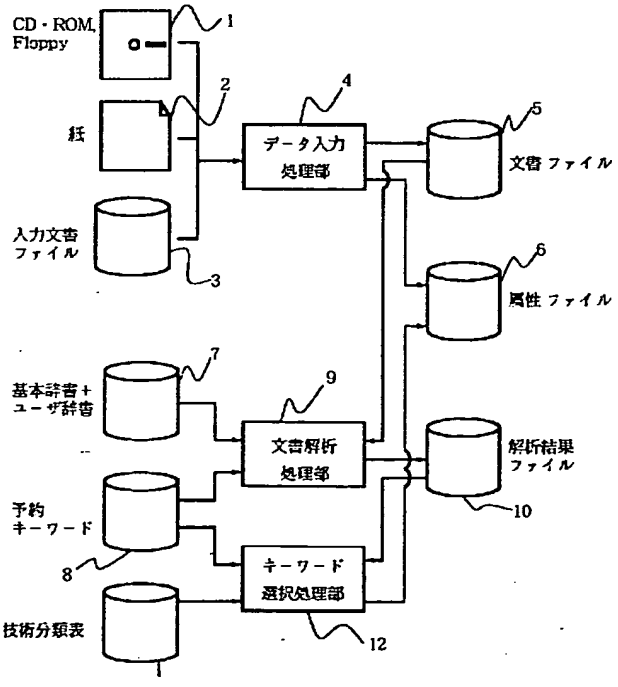
(74)代理人　弁理士　守山　辰雄

(54)【発明の名称】　キーワード付与装置

(57)【要約】

【目的】　ユーザの感覚に基づいたキーワードや文書データベースの種々な使用状況に対応したキーワードを文書に付与する。

【構成】　対象文書を走査して語を抽出する手段９と、抽出された語の所定の基準に従った重みを計数する手段９と、付与すべきキーワードの候補と該キーワード候補に対応する２以上の関連語の前記所定の基準に従った並びを含むキーワード付与基準データを保持する手段１１と、抽出された語の前記所定の基準に従った並びとキーワード付与基準データの関連語の並びが一致するかを照合する手段１２と、並びが一致すると判断された関連語に対応するキーワード候補を対象文書のキーワードとして付与する手段１２と、を備え、任意に設定したキーワードの候補を文書から抽出した語に基づいて当該文書のキーワードとして付与する。



【特許請求の範囲】
【請求項１】　任意に設定したキーワードの候補を文書から抽出した語に基づいて当該文書のキーワードとして付与するキーワード付与装置であって、
対象文書を走査して語を抽出する手段と、
前記抽出された語の所定の基準に従った重みを計数する手段と、
付与すべきキーワードの候補と該キーワード候補に対応する２以上の関連語の前記所定の基準に従った並びを含むキーワード付与基準データを保持する手段と、
前記抽出された語の前記所定の基準に従った並びと前記キーワード付与基準データの関連語の並びが一致するかを照合する手段と、
前記照合する手段によって並びが一致すると判断された関連語に対応するキーワード候補を前記対象文書のキーワードとして付与する手段と、
を備えたことを特徴とするキーワード付与装置。
【発明の詳細な説明】
【０００１】
【産業上の利用分野】本発明は、任意に設定したキーワードを文書から抽出した語に基づいて当該文書のキーワードとして付与するキーワード付与装置に関する。
【０００２】
【従来の技術】文書データベース等においては、蓄積された文書データの検索に利用するために、各文書にキーワードを付与することが行われている。キーワードの付与にはキーワードの的確性、検索に際しての利便性等が要求され、これらの要求を考慮して従来より種々の方式が提案されている。
【０００３】例えば、特開平４－２４３４７７号公報には、キーワードを予めユーザ辞書に登録しておき、このユーザ辞書を参照して文書に自然言語処理を行ってキーワードを抽出する方式が提案されている。また、特開平３－１３５６６９号公報には、検索の利便性を向上させるために、文書中から抽出した語に出現回数及び出現位置の重み付けをし、重み付けの度合いの高い語を重要キーワードとして抽出する方式が提案されている。
【０００４】また、特開平５－２５７９７９号公報には、キーワードの的確性を向上させるために、複数のキーワード候補をグループ化して予め登録しておき、文書中からキーワード候補を抽出した際に、このグループに属するキーワード候補も自動的に抽出する方式が提案されている。また、特開平２－２８７６９号公報には、キーワードの的確性を向上させるために、文書中に現れるキーワードの同義語や関連語からキーワードが表すキー概念を抽出し、キー概念の組合せを調べることにより文書全体の主題を表すキーワードを生成する方式が提案されている。
【０００５】
【発明が解決しようとする課題】しかしながら、上記した従来のキーワード付与方式にあっては、文書データベースの実際の使用状況を考察すると、機械的には生成し得ないユーザの感覚に基づいたキーワードや文書データベースの種々な使用状況に対応したキーワードを用いる方が便利な場合が多くある。例えば、技術分野の名称のように文書の内容全体を一言で表現したキーワード、担当者の名前のように或る範囲の文書を担当者毎に区別するキーワード、プロジェクトの名称のように或る範囲の文書を用途毎に区別するキーワード等のようにユーザが任意に設定したキーワードをデータベースに蓄積されている各文書に付与した方が、後のキーワード検索において便利な場合がある。また、キーワードによって分類化された各文書を更に細分類化するために、抽出されたキーワードから更に２次的なキーワードを抽出して各文書に付与した方が便利な場合もある。
【０００６】しかしながら、上記した従来のキーワード付与方式にあっては、文書に付与されるキーワードは、当該文書中から抽出されたキーワードから機械的に何らかの関連性があると判断された或る範囲の語でしかなく、ユーザが任意に設定した語にはなり得ないものであり、また、細分類化等のために抽出したキーワードから更に２次的なキーワードを抽出することもできなかった。
【０００７】本発明は、上記従来の事情に鑑みなされたもので、ユーザの感覚に基づいたキーワードや文書データベースの種々な使用状況に対応したキーワードを文書に付与することができるキーワード付与装置を提供することを目的とする。
【０００８】
【課題を解決するための手段】上記目的を達成するため、本発明のキーワード付与装置は、任意に設定したキーワード候補を文書から抽出した語に基づいて当該文書のキーワードとして付与するキーワード付与装置であって、対象文書を走査して語を抽出する手段と、前記抽出された語の所定の基準に従った重みを計数する手段と、付与すべきキーワードの候補と該キーワード候補に対応する２以上の関連語の前記所定の基準に従った並びを含むキーワード付与基準データを保持する手段と、前記抽出された語の前記所定の基準に従った並びと前記キーワード付与基準データの関連語の並びが一致するかを照合する手段と、前記照合する手段によって並びが一致すると判断された関連語に対応するキーワード候補を前記対象文書のキーワードとして付与する手段とを備えたことを特徴とする。
【０００９】ここに、上記の所定の基準とは、その語の出現頻度の他、その語が出現する文書中の位置で重み付けした出現頻度等であり、要は、その語の重要度を表す基準で、データベースの使用状況等に応じて任意に設定されるものである。

書から抽出された語の所定の基準に従った並びに基づいて、当該対象文書に付与するキーワードを決定するようにしている。すなわち、ユーザが設定した任意のキーワード候補に対応させて２以上の関連語を前記と同じ基準に従って並べておき、この関連語の並びと前記文書から抽出された語の並びが一致するキーワード候補を対象文書のキーワードとして採用する。

【００１１】したがって、文書から抽出した語から機械的には導き出せないユーザの感覚に基づいたキーワードや文書データベースの種々な使用状況に対応したキーワードをキーワード候補として設定し、これらキーワード候補の中から対象文書にキーワードを付与することができる。すなわち、技術分野の名称、担当者の名前、プロジェクトの名称等のように対象文書から機械的に導き出せない語もキーワードとして対象文書に付与することができ、また、文書を更に細分類化するための２次的なキーワードを対象文書に付与することができる。

【００１２】

【実施例】本発明の一実施例に係るキーワード付与装置を図面を参照して説明する。図１に示すように、本実施例のキーワード付与装置は、３つの形式の入力手段１、２、３と、入力された文書データを本装置で取り扱うことができるコードデータに変換するデータ入力処理部４と、コード変換された文書データを格納する文書ファイル手段５と、文書ファイル手段５に格納された文書の検索や分類に用いる情報を格納する属性ファイル手段６と、文書データを形態素解析するために用いる基本辞書及びユーザ辞書手段７と、形態素解析及びキーワード選択に用いる語を予約キーワードとして格納する予約キーワード格納手段８と、文書データを形態素解析する文書解析処理部９と、形態素解析されて構成語に分解された文書データを格納する解析結果ファイル手段１０と、文書に付与すべきキーワードの候補と該キーワード候補に対応する２以上の関連語の所定の基準に従った並びを記述した技術分類表（キーワード付与基準データ）を格納する技術分類表格納手段１１と、技術分類表に記述されたキーワード候補の中から所定の基準に該当する語をキーワードとして選択するキーワード選択処理部１２とを備えている。

【００１３】入力手段１からはフロッピーディスクやＣＤ－ＲＯＭ等からの文書のコード情報やイメージ情報が入力され、入力手段２からは印刷物等から読み取った文書のイメージ情報が入力され、入力手段３からは磁気ディスクに格納された文書データやネットワークで接続された他のシステムからの文書データが入力される。

【００１４】データ入力処理部４は、入力手段１、２、３から入力された文書データを、イメージデータはコードデータに変換し、更にコードデータも本装置で取り扱た、データ入力処理部４は、この処理に際して文書名、文書の作成年月日、文書の作成者等といった格納された文書の検索や分類に便利な情報が得られる場合には、これら情報を属性ファイル手段６に格納する。属性ファイル手段６にはこのような情報の他、キーワード選択処理部１２で選択されたキーワードも入力され、属性ファイル手段６はこれら情報及びキーワードを文書毎に対応付けて格納する。

【００１５】基本辞書及びユーザ辞書手段７並びに予約キーワード格納手段８の内容は文書解析処理部９による文書の形態素解析に用いられ、特に、予約キーワード格納手段８には技術分類表格納手段１１の関連語と一致することを意図した多数の語が格納されている。これら辞書手段７並びに予約キーワード格納手段８の内容は最適な形態素解析と語の抽出が行われるように、ユーザーによって予め作成及び編集されている。文書解析処理部９は、辞書手段７並びに予約キーワード格納手段８を参照して、文書ファイル手段５に格納された文書データを文書毎に形態素解析し、この解析結果を文書毎に解析結果ファイル手段１０に格納する。

【００１６】技術分類表格納手段１１はキーワード付与基準データとしての技術分類表を格納しており、この技術分類表には文書に付与すべき多数のキーワード、および、これらキーワードにそれぞれ対応した複数の関連語が記述されている。すなわち、技術分類表は図２に示すような構成となっており、技術分類の欄には文書に付与すべきキーワードとして「ファイル検索」、「ファイル管理、」ディスク接続」等といった語が記述され、重要度順位の欄には各語に対応させて「ファイル、検索、・・・」、「ファイル、管理、・・・」、「Ｄｉｓｋ、制御、・・・」等といった複数の関連語が記述されている。

【００１７】本実施例では所定の基準として語の文書中における出現頻度を採用しており、各重要度順位の欄における関連語はこの出現頻度の高い順に左から配列されている。例えば、付与すべきキーワード「ファイル検索」に対応して関連語「ファイル、検索、・・・」が出現頻度の高い順に並べて記述されている。この技術分類表の技術分類の欄及び重要度順位の欄に記述される語はユーザによって任意に設定されるのもであり、本実施例では、文書を技術分野毎に分類することを意図しているため、文書に付与すべきキーワードを記述分野を示す語として技術分類の欄に記述してある。

【００１８】キーワード選択処理部１２は解析結果ファイル手段１０に格納されている構成語に分解された文書データを文書毎に処理するものであり、この文書データ中から予約キーワード格納手段８に登録されている語を抽出し、抽出された語の当該文書中における出現回数を

関連の語並びに照合し、抽出された語の並びが関連語の並びに一致する技術分類の欄の語を文書のキーワードとして選択して、属性ファイル手段６に当該文書のキーワードとして登録する。

【００１９】上記構成のキーワード付与装置における処理を図３乃至図６に示すフローチャートを参照して説明する。まず、キーワード付与装置への文書データの入力処理は、データ入力処理部４によって図３に示す手順で行われる。すなわち、入力手段１、２、３から入力された文書データを読み取り（ステップＳ１）、この文書データを必要なコード変換を行って文書ファイル手段５に格納するとともに（ステップＳ２）、文書データから抽出した文書名等の属性データを属性ファイル手段６に格納する（ステップＳ３）。この属性ファイル手段６には文書毎に属性ファイルが形成され、各文書に対応して属性データが格納される。

【００２０】そして、入力手段１、２、３からは各手段に入力された文書数を示す信号や入力された全ての文書の出力終了を示す信号がデータ入力処理部４に入力されており、データ入力処理部４はこの信号に基づいて全ての文書データの入力処理が終了したかを判断し（ステップＳ４）、全ての文書データの入力処理が終了するまで上記ステップＳ１乃至Ｓ３の処理を繰り返し行う。

【００２１】次いで、上記のようにして入力された文書データの解析処理は、文書解析処理部９によって図４に示す手順で行われる。すなわち、文書ファイル手段５から１つの文書のデータを読み取り（ステップＳ１１）、基本辞書及びユーザ辞書手段７並びに予約キーワード格納手段８の内容を参照して、この文書データを走査しつつ形態素解析して構成語に分解し（ステップＳ１２）、構成語に分解された文書データを解析結果ファイル１０に格納する（ステップＳ１３）。この分解された文書データは構成語を区切コードによって区切った構造であり、各文書毎に対応付けたファイルとして解析結果ファイル１０に格納される。

【００２２】そして、上記データ入力処理部４で入力した文書数と実際に解析処理が終了した文書数とから、文書ファイル手段５に格納された全ての文書データについての解析処理が終了したかを判断し（ステップＳ１４）、全ての文書データの解析処理が終了するまで上記ステップＳ１１乃至Ｓ１３の処理を繰り返し行う。

【００２３】次いで、上記のようにして解析された文書データに基づいたキーワードの選択処理は、キーワード選択処理部１２によって図５及び図６に示す手順で行われる。すなわち、解析結果ファイル手段１０から１つの文書のデータを読み取り（ステップＳ２１）、予約キーワード格納手段８に格納されている語に一致する語をこの文書データから抽出して、抽出された語の当該文書データとを確認した語（ステップＳ２３）、当該抽出された語をキーワードとして属性ファイル手段６の対応する属性ファイルに格納する（ステップＳ２４）。なお、この抽出された語によるキーワードは従来より一般的なキーワードであり、本発明に特有なキーワードは後述するステップＳ３４で登録されるキーワードである。

【００２４】そして、上記のような抽出語のキーワード登録が終了すると、予約キーワード格納手段８に格納されている語の数に基づいて、予約キーワード格納手段８に格納されている全ての語について文書データからの抽出処理が終了したかを判断し（ステップＳ２５）、終了していない場合には上記ステップＳ２２乃至Ｓ２４の処理を繰り返し行う。一方、全ての予約キーワード語について文書データからの抽出処理が終了した場合には、抽出した各語の出現頻度順の並びを技術分類表格納手段１１の格納された技術分類表の内容に参照する処理を行う。

【００２５】まず、テーブル（１）に抽出した各語を出現頻度順に並べ（ステップＳ２６）、図２に示した技術分類表の技術分類の欄を順次示すポインタ（１）を最上段の欄（「ファイル検索」）にセットする（ステップＳ２７）。次いで、テーブル（１）のポインタ（２）を最上段の抽出語（すなわち、出現頻度の最も高い語）にセットし（ステップＳ２８）、技術分類表のポインタ（１）が示す技術分類の欄に対応する重要度順位の欄の関連語を順次示すポインタ（３）を先頭の（すなわち、出現頻度が大きいとした）関連語（「ファイル」）にセットする（ステップＳ２９）。次いで、ポインタ（２）の示す抽出語とポインタ（３）が示す関連語とを照合し（ステップＳ３０）、両者が一致するかを判断する（ステップＳ３１）。すなわち、この場合には、対象の文書データから最も出現頻度が高いとして抽出された語が、図２の技術分類表の最上段の欄の「ファイル」という関連語に一致するかを判断する。

【００２６】この判断の結果、一致する場合には、ポインタ（２）を１つ増加させてテーブル（１）の次の抽出語を示させるとともに、ポインタ（３）も１つ増加させて技術分類表のポインタ（１）が示す技術分類の欄に対応する重要度順位の欄の次の関連語（「検索」）を示させる（ステップＳ３２）。この後、ポインタ（２）又はポインタ（３）の先が終了か、すなわち、増加されたポインタ（２）がテーブル（１）の最後の抽出語を示している又は増加されたポインタ（３）が技術分類表のポインタ（１）が示す技術分類の欄に対応する重要度順位の欄の最後の関連語を示しているかを判断する（ステップＳ３３）。すなわち、技術分類表のポインタ（１）が示す技術分類の欄に対応する重要度順位の欄の全ての関連語について上記の照合が終了したかを判断する。

は、上記のステップＳ３０以降の処理を繰り返し行う一方、上記の照合による一致がとれつつ重要度順位の欄の全ての関連語について上記の照合が終了している場合には、技術分類表のポインタ（１）が示す技術分類の語を対象の文書データのキーワードとして属性ファイル６に登録する（ステップＳ３４）。すなわち、技術分類表の重要度順位の欄に出現頻度を基準として記述された関連語の並びに、対象の文書データから抽出された語の出現頻度順の並びとが一致した場合に、この重要度順位の欄に対応する技術分類の欄の語がキーワードとして登録される。例えば、対象の文書データから「ファイル」、「検索」、・・・が頻度の高い順に抽出された場合には、対応する「ファイル検索」という技術分類を表す語が対象文書の２次的なキーワードとして登録される。

【００２８】一方、上記のステップＳ３１で一致がとれない場合、又は、上記ステップＳ３４による登録がなされた場合には、ポインタ（１）を１つ増加させて技術分類表の次の段についての照合を行う（ステップＳ３５）。すなわち、現在のポインタ（１）が既に技術分類表の最下段にまで達していないことを確認した後（ステップＳ３６）、技術分類表の「ファイル管理」の欄、「ディスク管理」の欄、・・・について順次上記のステップＳ２８乃至Ｓ３５の処理を繰り返し行う。この結果、対象としている文書データについて、抽出語の並びと関連語の並びの一致がある場合には、技術分類表の技術分類の欄に記述してある語が２次的なキーワードとして付与される。

【００２９】上記した１つの文書データについてのキーワード選択処理が終了すると、解析結果ファイル手段１０に格納されている全ての文書データについて上記の処理が終了したかを判断し（ステップＳ３７）、終了していない場合には解析結果ファイル手段１０に格納されている次の文書データについて上記のステップＳ２１乃至Ｓ３６の処理を繰り返し行って、各文書データに技術分類表から２次的なキーワードを選択して付与し、属性ファイル手段６に登録する。

【００３０】ここで、例えば、文書から抽出された語が、「ファイル」、「検索」、「管理」の３つであったとし、或る文書Ａでは「ファイル」が５回、「検索」が３回、「管理」が１回使われており、他の文書Ｂでは「ファイル」が５回、「検索」が１回、「管理」が３回使われていたとすると、従来の方式でキーワードを付与すると、文書Ａ、Ｂ共に同じキーワードが付与されることとなる。これに対し、本実施例のキーワード付与装置によると、文書Ａには上記３つの１次的なキーワードの他に「ファイル検索」という２次的なキーワードが付与され、文書Ｂには上記３つの１次的なキーワードの他に「ファイル管理」という２次的なキーワードが付与されによると、文書に付与される１次的キーワード（抽出語）の出現頻度によって選択される２次的なキーワード（技術分類）も当該文書に付与されることとなる。このため、１次的なキーワードから機械的には導き出せないユーザの感覚に基づいた２次的なキーワードや文書データベースの種々な使用状況に対応した２次的なキーワードを１次的キーワードの重要度を加味して文書に付与することができ、文書の細分類化や後のキーワード検索において多大な利便性を実現することができる。

【００３２】次に、本発明の他の一実施例に係るキーワード付与装置を図面を参照して説明する。なお、前記の実施例と同一部分には同一符号を付して重複する説明は省略する。本実施例のキーワード付与装置は、文書に対して技術分類表から自動的に付与された２次的なキーワード（技術分類）を利用して、当該文書から抽出された語を１次的なキーワードとして登録するかの判断を担当のユーザからの会話形式の指示で行うものである。

【００３３】本実施例のキーワード付与装置は、図７に示すように、前記の実施例と異なる構成として、キーワード選択処理部１２の代わりにキーワード候補選択部２１を備える他、新たな構成として、抽出キーワードインデックスファイル手段２２と、キーワード設定処理部２３と、表示装置２４を有した指示装置２５とを備えている。

【００３４】キーワード候補選択部２１は、前記のキーワード選択処理部１２と同様に、解析結果ファイル手段１０に格納されている解析された文書データから予約キーワード格納手段８に格納されている語と一致する語を抽出し、その抽出語の当該文書における出現回数をカウントし、これら抽出語を出現頻度順に並べて技術分類表格納手段１１に格納されている技術分類表（図２参照）に照合し、関連語の並びが一致する技術分類の欄の語を判別する。そして、このキーワード候補選択部２１は、対象としている文書データの文書識別子（文書ＩＤ）及び判別した技術分類の欄の語と共に、当該文書データから抽出した語を１次的キーワードの候補として抽出キーワードインデックスファイル手段２２に格納する。

【００３５】抽出キーワードインデックスファイル手段２２には図８に示すようなインデックスファイルが格納されており、上記の文書ＩＤ及び技術分類の欄の語に対応して、１次的キーワードの候補である抽出語が記述されている。例えば、本実施例では、文書ＩＤ「文書Ｔ」と「文書Ａ」に対応して抽出語「ファイル」、「検索」及び２次的キーワード候補の技術分類として「ファイル検索」が記述され、また、文書ＩＤ「文書Ｘ」に対応して抽出語「ファイル」、「管理」及び２次的キーワード候補の技術分類として「ファイル管理」が記述される。

【００３６】指示装置２５はユーザが担当する技術分類

ワードインデックスファイル手段２２のインデックスファイルを検索し、当該技術分類に対応する文書ＩＤ及び抽出語を探索する。そして、キーワード設定処理部２３は、探索された文書ＩＤに該当する文書データを解析結果ファイル手段１０から読み出して文書を表示装置２４に表示すると共に、表示した文書中で探索された抽出語を他の語とは色を変える、或いは、ハイライトする等して表示する。そして、指示装置２５から表示された抽出語を登録する指示が入力された場合には、この抽出語を表示されている文書のキーワードとして属性ファイル手段６に登録する。なお、この抽出語のキーワード登録と共に、対応する技術分対の語も２次的キーワードとして属性ファイル手段６に登録してもよい。

【００３７】上記構成のキーワード付与装置における処理を図９乃至図１０に示すフローチャートを参照して説明する。なお、キーワード付与装置への文書データの入力処理と文書データの解析処理は前記の実施例と同様であるので説明を省略する。まず、文書データからキーワード候補を抽出してインデックスファイルを作成する処理は、キーワード候補選択処理部２１によって図９に示す手順で行われる。すなわち、解析結果ファイル手段１０から１つの文書データを読み取り（ステップＳ４１）、予約キーワード格納手段８に格納されている語に一致する語をこの文書データから抽出して、抽出された語の当該文書データ中における出現回数をカウントする（ステップＳ４２）。そして、予約キーワード格納手段８に格納されている語の数に基づいて、予約キーワード格納手段８に格納されている全ての語について文書データからの抽出処理が終了したかを判断し（ステップＳ４３）、終了していない場合には上記ステップＳ４２の処理を繰り返し行う。

【００３８】一方、全ての予約キーワード語について文書データからの抽出処理が終了した場合には、前記の実施例と同様にして、抽出した各語の出現頻度順の並びを技術分類表格納手段１１の格納された技術分類表の内容に順次参照し、両者の並びが一致する技術分類の語を検索する（ステップＳ４４、Ｓ４５、Ｓ４６）。この結果、一致を見い出せた場合には、その技術分類の語を抽出キーワードインデックスファイル手段２２に格納されているインデックスファイルの技術分類の欄に登録し（ステップＳ４７）、技術分類表の全項目を検索しても一致が見い出せなかった場合には、インデックスファイルの技術分類の欄に「その他」を登録する（ステップＳ４８）。

【００３９】次いで、この技術分類の欄に対応させて、上記の抽出語を全てインデックスファイルの抽出語の欄に登録し（ステップＳ４９）、更に、上記の文書データの文書ＩＤをインデックスファイルの文書ＩＤの欄に登記の処理が終了したかを判断し（ステップＳ５１）、終了していない場合には解析結果ファイル手段１０に格納されている次の文書データについて上記のステップＳ４１乃至Ｓ５０の処理を繰り返し行って、各文書ＩＤに対応したインデックスファイルを作成する。

【００４０】次に、上記のようにして作成されたインデックスファイルに基づいたキーワードの設定処理は、キーワード設定処理部２３によって図１０に示す手順で行われる。すなわち、ユーザが自分の担当する技術分類の指定を指示装置２５から入力すると（ステップＳ６１）、この技術分類に該当する項目をインデックスファイルから１つ読み出し（ステップＳ６２）、その項目に記述されている文書ＩＤに基づいて解析結果ファイル手段１０から対応する文書データを読み出す（ステップＳ６３）。

【００４１】次いで、インデックスファイルの対応する抽出語の欄から抽出語を１つ読み出して上記の文書データに照合し、文書データ中の一致する語を他の語とは色を変え、当該文書データによる文書を表示装置２４に表示する（ステップＳ６４）。そして、色を変えて表示された抽出語をキーワードとして登録する指示がユーザによって指示装置２５から入力されたかを判断し（ステップＳ６５）、この指示が入力された場合には、この抽出語を表示されている文書のキーワードとして属性ファイル手段６に登録する。

【００４２】一方、登録する指示がなかった場合、或いは、上記の登録が終了した後には、インデックスファイルの対応する抽出語の欄に記述された全ての抽出語について登録するか否かの指示を受けたかを判断し（ステップＳ６７）、抽出語の全てについての指示を受けていない場合には、対応する抽出語の欄に記述された次の抽出語について上記のステップＳ６４乃至Ｓ６６の処理を繰り返し行う。そして、対応する全ての抽出語について登録するか否かの指示を受けて、これら抽出語についての処理が終了した場合には、ユーザから入力された技術分類についてインデックスファイルに登録されている全ての項目の処理が終了したかを、インデックスファイルの登録件数から判断する（ステップＳ６８）。この結果、終了していない場合には、入力された技術分類のインデックスファイル中の次の項目（すなわち、次の文書）について上記のステップＳ６２以降の処理を繰り返し行う。

【００４３】例えば、ユーザから「ファイル検索」という技術分類が入力されると、図８に示すインデックスファイルに基づいて、文書Ｔが表示装置に表示され、当該文書Ｔ中の「ファイル」という抽出語がキーワード候補として色を変えて表示される。そして、ユーザからキーワードとして登録する指示があった場合には、この抽出

同様な処理がなされ、ユーザによる指示に基づいてキーワードとして登録される。このように、インデックスファイルの文書Ｔの項目についての処理が終了すると、次の文書Ａの項目についても上記と同等な処理がなされ、「ファイル」、「検索」という抽出語がユーザの指示に基づいてキーワードとして登録される。

【００４４】すなわち、本実施例のキーワード付与装置によると、文書から抽出した語を当該文書のキーワードとして登録するか否かの判断を、技術分類表に登録してある２次的なキーワードに基づいて割り当てた専門的なユーザに担当させることができ、正確なキーワードの付与を効率よく行うことができる。

【００４５】上記した各実施例ではキーワード付与基準データを技術分類表として２次的キーワードの候補を技術分野を表す語としたため、例えば、特許公報の調査のように技術分野毎に専門的な知識を必要とする作業において、最適な担当者を割り当てて、データベースを用いた効率的な作業を実現することができる。なお、本発明はこれに限らず、図１１に示すように２次的キーワードの候補を文書を担当することを割り当てられた担当者名とする等、ユーザの感覚に基づいた最適な語や文書データベースの種々な使用状況に対応した語とすることができる。

【００４６】また、上記した各実施例では抽出語及び関連語の並びと決める基準を文書中での出現頻度としたが、本発明では必要に応じて種々な基準を適用することができる。例えば、文書中での語の出現位置で重み付けすれば、特許公報の調査においては特許請求の範囲の語は重くその他は軽くして、より使用状況やユーザーの感覚にあったキーワードを各特許公報に付与することができる。

【００４７】また、抽出語の並びと関連語の並びとを照合方法には、上記の各実施例で示したように全ての語の並びが一致する場合のみならず、本発明では一部の語の並びが一致する場合でも２次的キーワードの登録をするようにしてもよい。例えば、抽出語の並びがＡ、Ｂ、Ｃ、Ｄで、関連語の並びがＡ、Ｂ、Ｄあった場合に、Ａ、Ｂ、Ｄという一部の語の並びから、これらが一致するものとして処理したり、或いは、抽出語の並びがＡ、Ｂ、Ｃ、Ｄで、関連語の並びがＡ、Ｂ、Ｃ、Ｅ、Ｄあった場合に、Ａ、Ｂ、Ｃ、Ｄという一部の語の並びから、これらが一致するものとして処理するようにしてもよい。

【００４８】

【発明の効果】以上詳細に説明したように、本発明のキーワード付与装置によると、任意に設定したキーワードの候補と該キーワード候補に対応する２以上の関連語をキーワード付与基準データとして保持しておき、対象文書から抽出された語の並びと関連語の並びが一致するかを照合し、これらの並びが一致する場合に関連語に対応するキーワード候補を対象文書のキーワードとして付与するようにしたため、文書から抽出した語から機械的には導き出せないユーザの感覚に基づいた語や文書データベースの種々な使用状況に対応した語を対象文書に２次的なキーワードとして付与することができる。このため、データベースとして蓄積された文書のキーワード検索を効率的に行うことができ、また、各文書を細分類化して能率的な文書データベースを構築することができるという効果を奏する。

【図面の簡単な説明】

【図１】本発明の一実施例に係るキーワード付与装置の構成を示すブロック図である。

【図２】本発明の一実施例に係る技術分類表の構成を示す概念図である。

【図３】本発明の一実施例に係るデータ入力処理の手順を示すフローチャートである。

【図４】本発明の一実施例に係る文書解析処理の手順を示すフローチャートである。

【図５】本発明の一実施例に係るキーワード選択処理の手順を示すフローチャートである。

【図６】本発明の一実施例に係るキーワード選択処理の手順を示すフローチャートである。

【図７】本発明の他の一実施例に係るキーワード付与装置の構成を示すブロック図である。

【図８】本発明の他の一実施例に係るインデックスファイルの構成を示す概念図である。

【図９】本発明の他の一実施例に係るキーワード選択候補処理の手順を示すフローチャートである。

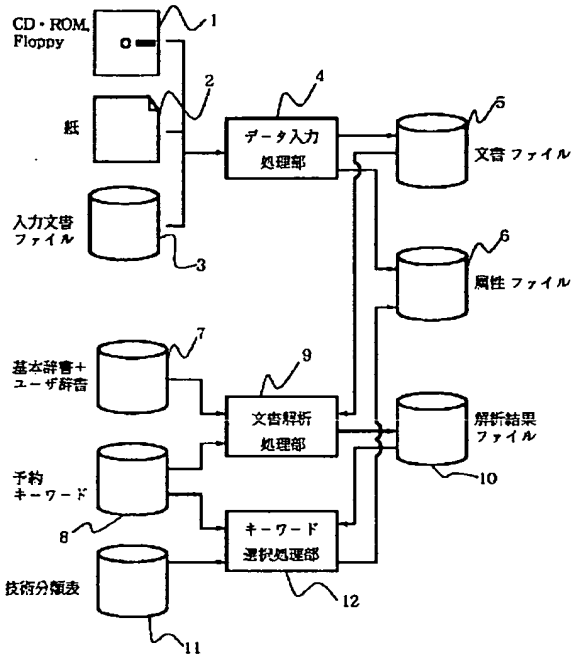
【図１０】本発明の他の一実施例に係るキーワード設定処理の手順を示すフローチャートである。

【図１１】技術分類表の他の一例の構成を示す概念図である。

【符号の説明】

４　データ入力処理部
５　文書ファイル手段
６　属性ファイル手段
８　予約キーワード格納手段
９　文書解析処理部
１０　解析結果ファイル手段
１１　技術分類表格納手段（キーワード付与基準データ）
１２　キーワード選択処理部

【図１】



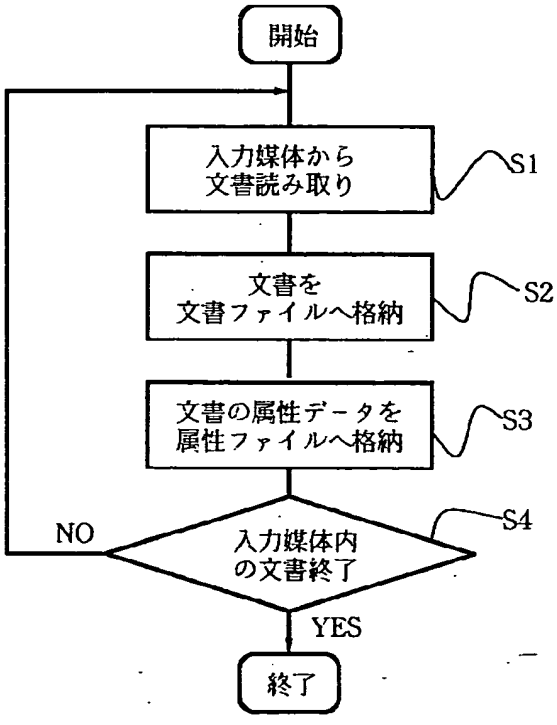
【図２】

技術分類	重要度順位
ファイル検索	ファイル、検索、・・・・
ファイル管理	ファイル、管理、・・・・
ディスク装置	Disk、制御、・・・・・

【図１１】

担当者	重要度順位
山口義昭	ファイル、検索
富士太郎	ファイル、管理
赤坂次郎	Disk、制御
⋮	⋮

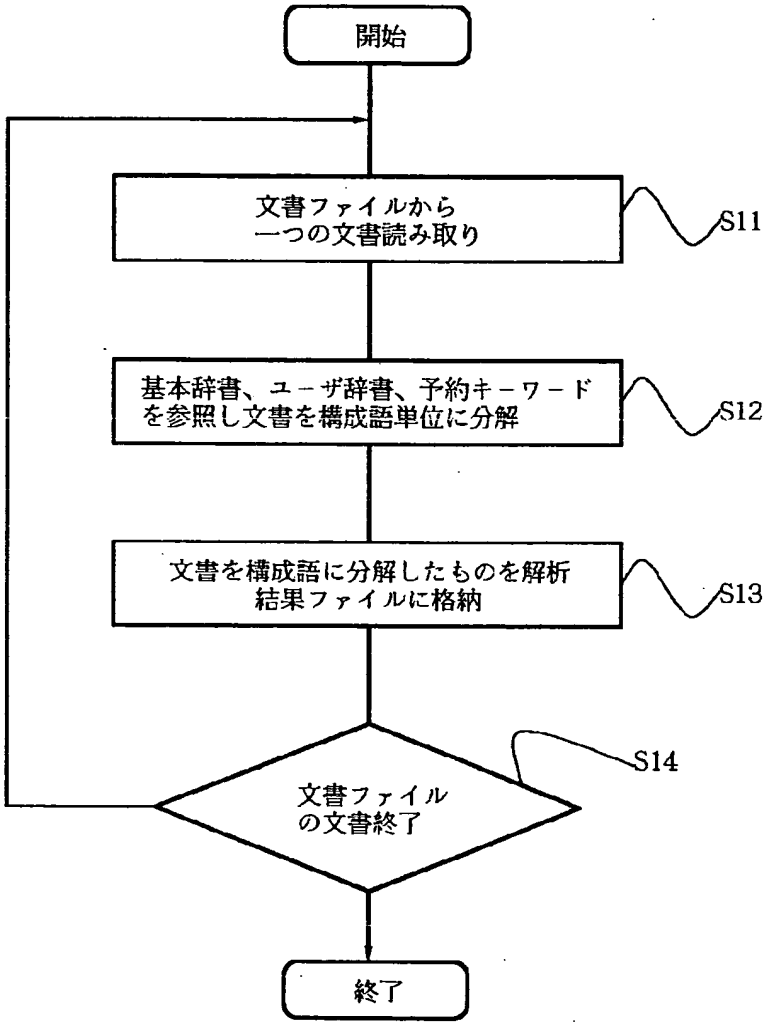
【図３】



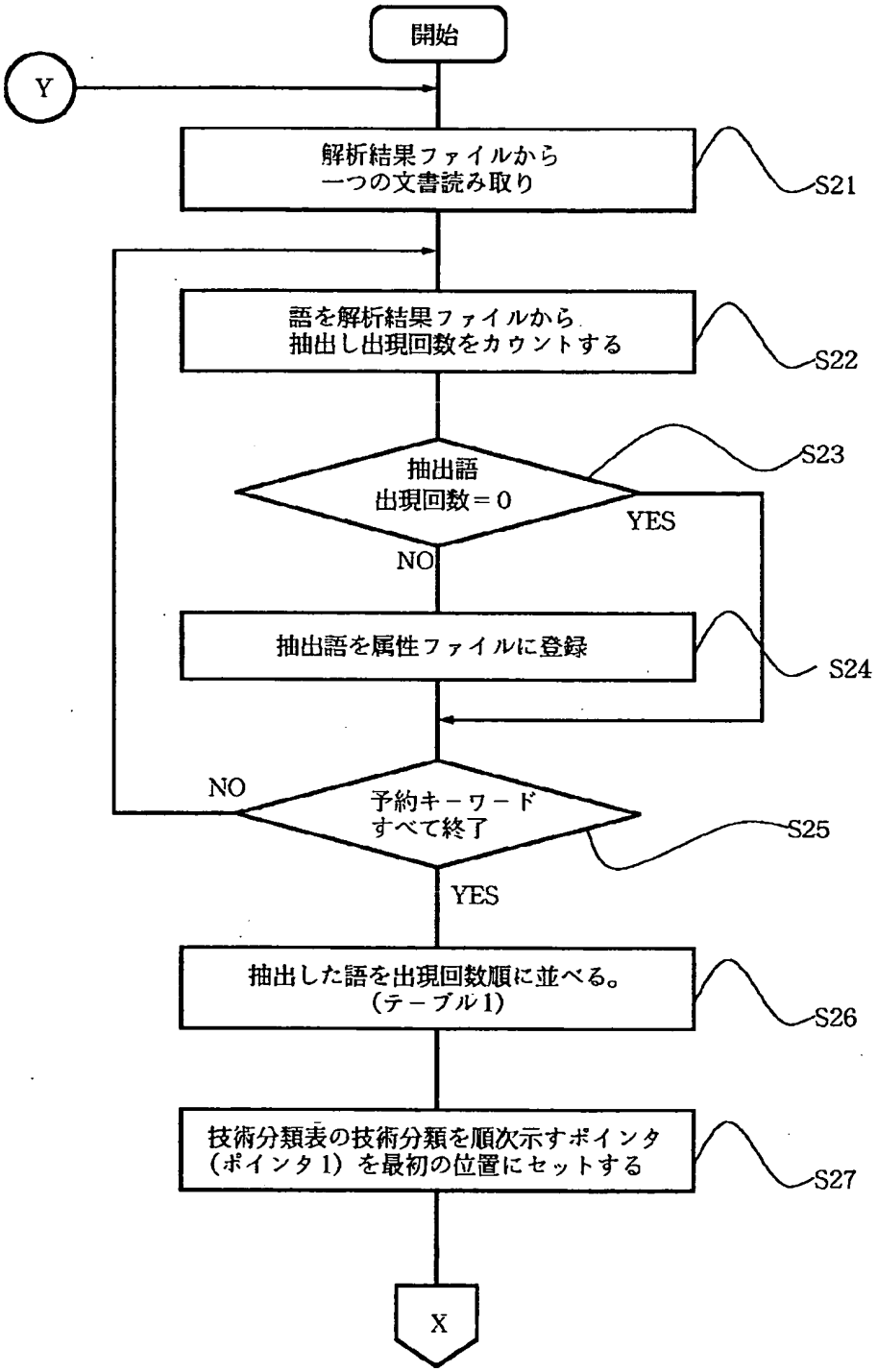
【図８】

技術分類	抽出語	文書ID
ファイル検索	ファイル、検索	文書T
ファイル検索	ファイル、検索	文書A
⋮	⋮	⋮
ファイル管理	ファイル、管理	文書X
⋮	⋮	⋮
ディスク装置	Disk、制御	文書H

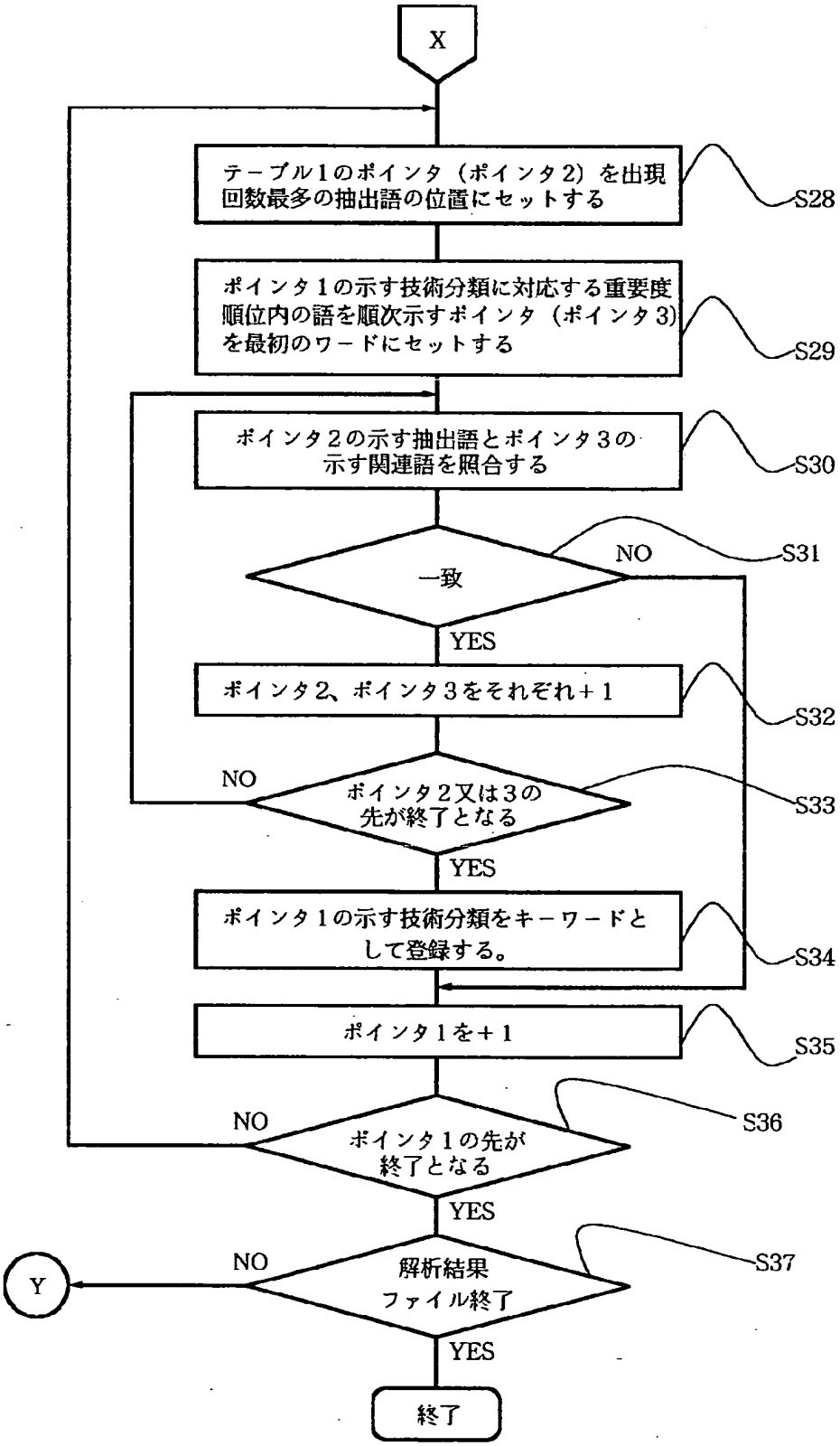
【図４】


開始
文書ファイルから
一つの文書読み取り
S11
基本辞書、ユーザ辞書、予約キーワード
を参照し文書を構成語単位に分解
S12
文書を構成語に分解したものを解析
結果ファイルに格納
S13
文書ファイル
の文書終了
S14
終了

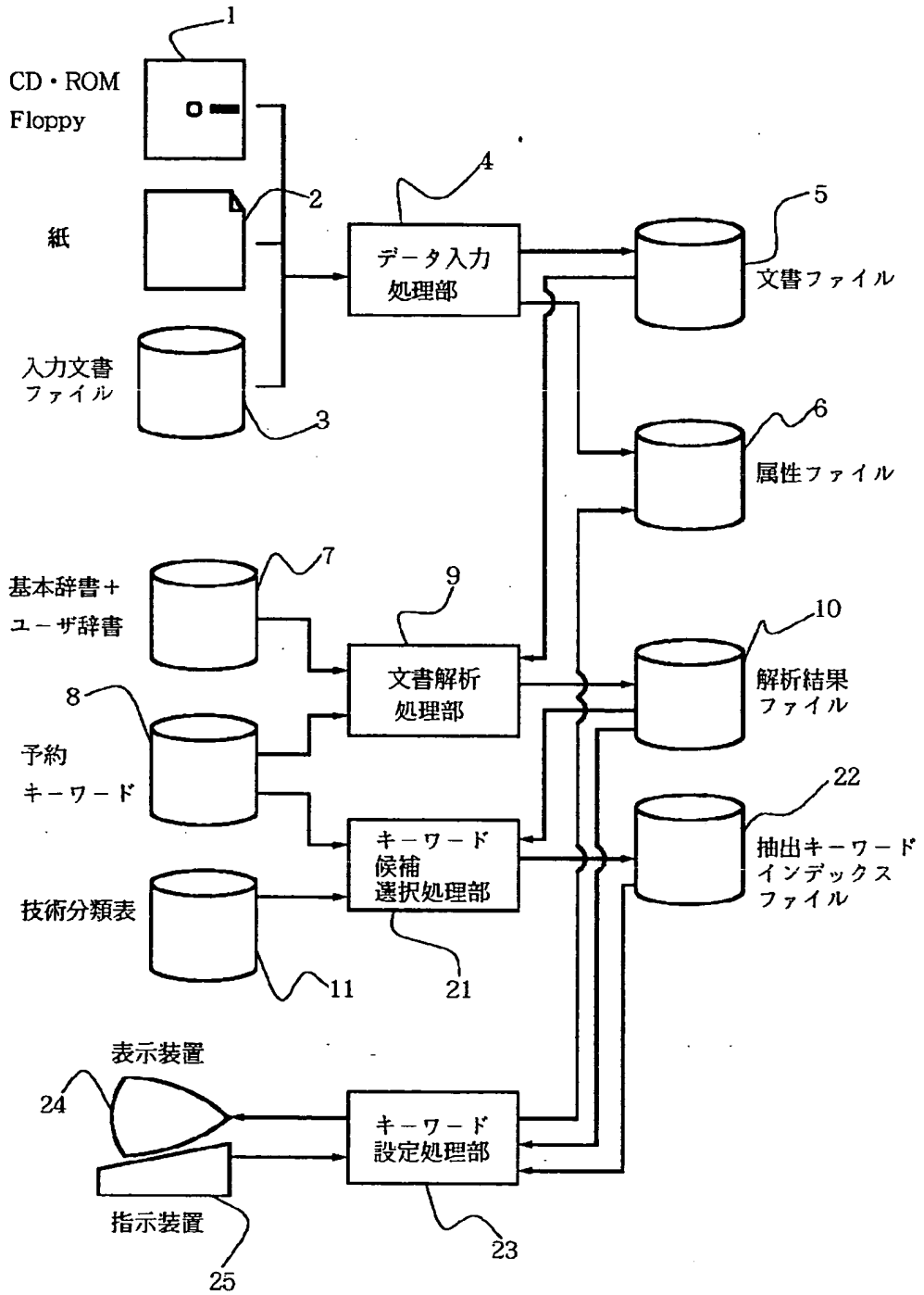
【図５】


開始
Y
解析結果ファイルから
一つの文書読み取り
S21
語を解析結果ファイルから
抽出し出現回数をカウントする
S22
抽出語
出現回数＝0
S23
YES
NO
抽出語を属性ファイルに登録
S24
NO
予約キーワード
すべて終了
S25
YES
抽出した語を出現回数順に並べる。
（テーブル1）
S26
技術分類表の技術分類を順次示すポインタ
（ポインタ1）を最初の位置にセットする
S27
X

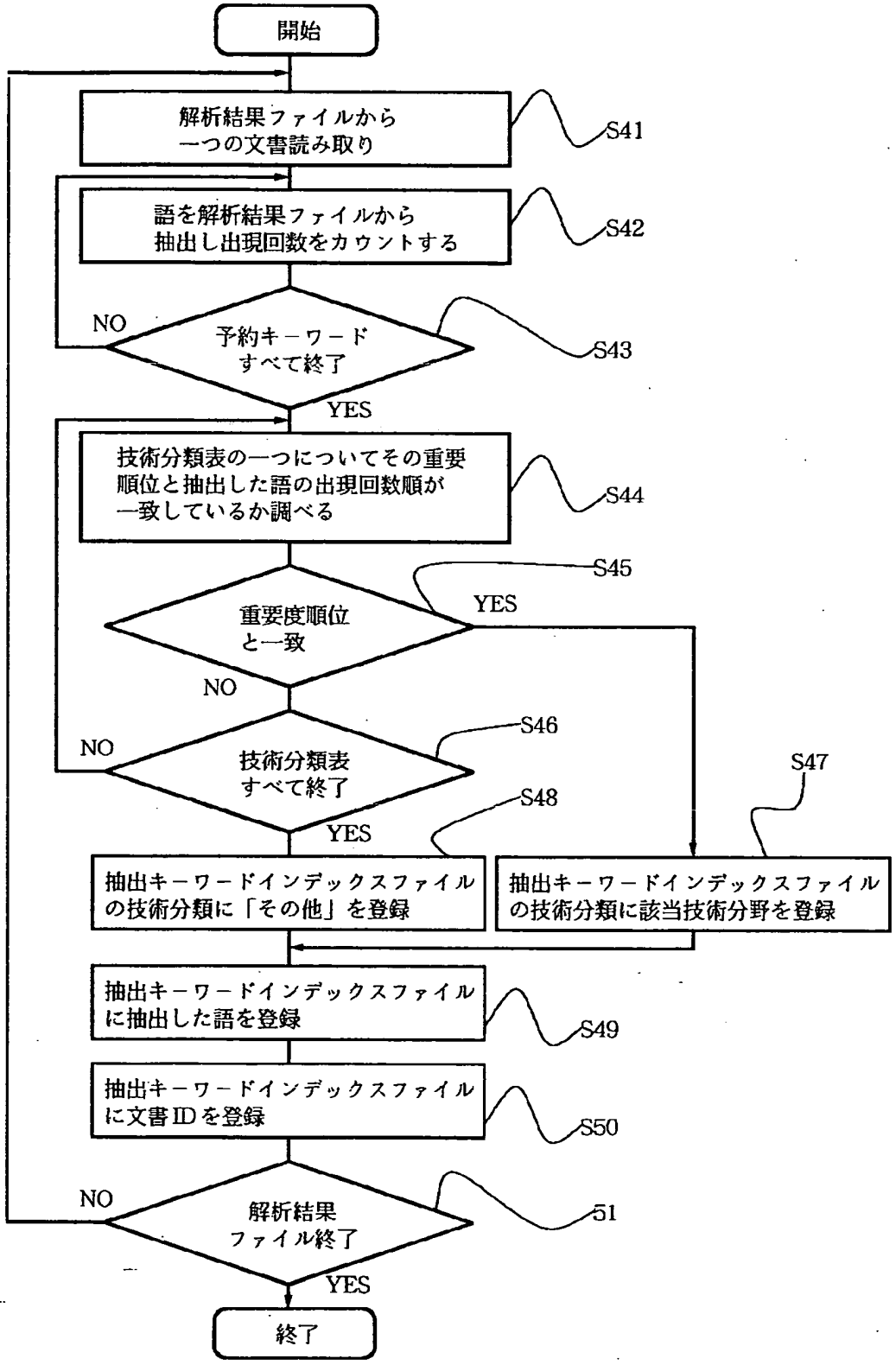
【図６】


X
テーブル１のポインタ（ポインタ２）を出現回数最多の抽出語の位置にセットする
S28
ポインタ１の示す技術分類に対応する重要度順位内の語を順次示すポインタ（ポインタ３）を最初のワードにセットする
S29
ポインタ２の示す抽出語とポインタ３の示す関連語を照合する
S30
一致
NO
S31
YES
ポインタ２、ポインタ３をそれぞれ＋１
S32
NO
ポインタ２又は３の先が終了となる
S33
YES
ポインタ１の示す技術分類をキーワードとして登録する。
S34
ポインタ１を＋１
S35
NO
ポインタ１の先が終了となる
S36
YES
NO
解析結果ファイル終了
S37
Y
YES
終了

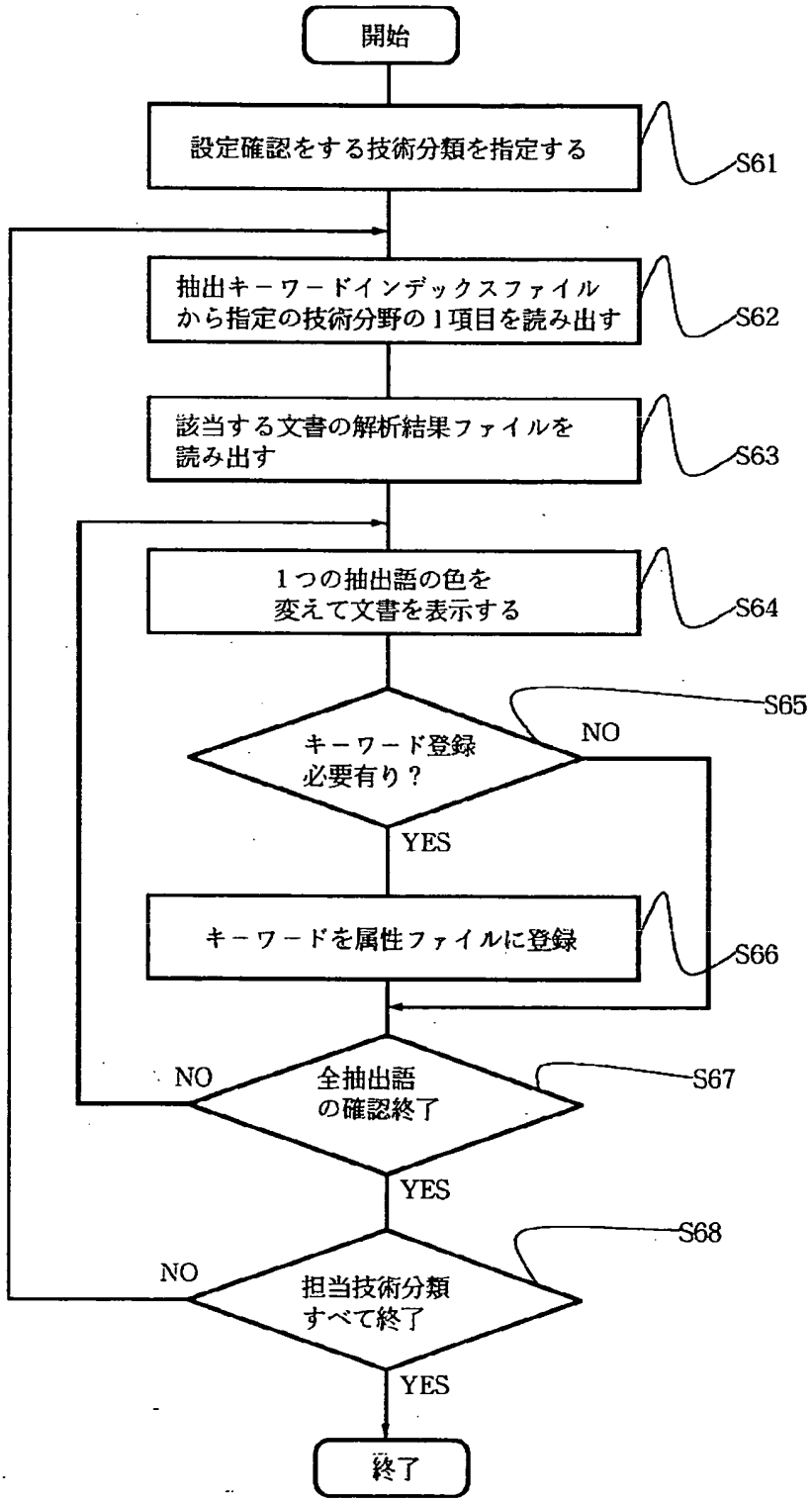
【図７】


1
CD・ROM
Floppy
2
紙
入力文書
ファイル
3
4
データ入力
処理部
5
文書ファイル
6
属性ファイル
基本辞書＋
ユーザ辞書
7
8
予約
キーワード
9
文書解析
処理部
10
解析結果
ファイル
22
抽出キーワード
インデックス
ファイル
キーワード
候補
選択処理部
技術分類表
11
21
表示装置
24
キーワード
設定処理部
指示装置
25
23

【図９】


開始
解析結果ファイルから
一つの文書読み取り
S41
語を解析結果ファイルから
抽出し出現回数をカウントする
S42
NO
予約キーワード
すべて終了
S43
YES
技術分類表の一つについてその重要
順位と抽出した語の出現回数順が
一致しているか調べる
S44
S45
重要度順位
と一致
YES
NO
S46
NO
技術分類表
すべて終了
S47
S48
YES
抽出キーワードインデックスファイル
の技術分類に「その他」を登録
抽出キーワードインデックスファイル
の技術分類に該当技術分野を登録
抽出キーワードインデックスファイル
に抽出した語を登録
S49
抽出キーワードインデックスファイル
に文書IDを登録
S50
NO
解析結果
ファイル終了
51
YES
終了

【図１０】


開始
設定確認をする技術分類を指定する
S61
抽出キーワードインデックスファイル
から指定の技術分野の１項目を読み出す
S62
該当する文書の解析結果ファイルを
読み出す
S63
１つの抽出語の色を
変えて文書を表示する
S64
S65
キーワード登録
必要有り？
NO
YES
キーワードを属性ファイルに登録
S66
NO
全抽出語
の確認終了
S67
YES
S68
NO
担当技術分類
すべて終了
YES
終了